DOI 10.34658/9788366741102.5

Search for outliers by fuzzy logic systems – general concepts

Marcin Kacprowicz[0000−0003−4381−1359]

Lodz University of Technology
Institute of Information Technology
Wólczańska 215, 90-924 Łódź, Poland
marcin.kacprowicz@p.lodz.pl

Abstract. *This article presents both a review of the literature and research proposals on the use of artificial integration, in particular fuzzy logic and fuzzy logic systems to search for and find exceptions outlier.*
Keywords: *artificial intelligence, search for outliers, finding outliers, fuzzy logic*

1. Introduction

Analyzing large data sets is an impossible task without using machines. However, the analysis with the use of computers in a classic approach boils down to the application of statistical methods, which do not always allow obtaining information useful from the point of view of the final recipient. Finding data that "sticks" from others is a particularly difficult task. We define such data as outliers, whose definitions can be found in many literature sources[1, 2]. Figure 1 shows an example of an exception, but it is only a simplified form. Most generally it can be said that the outlier is a data which in its construction differs from the other data in the set to the extent that it indicates the fact that it is generated by a different mechanism than the other data. The data outliers problem can be considered in two ways. The first of these is the matter of looking for exceptions, i.e. finding the existence of exceptions in the data. The second is finding exceptions, i.e. indicating a specific data which is a distance from the others and a simultaneous indication of how it differs from the others. The use of such methods has a very wide range, from activities at the level of the entertainment industry, e.g. computer games, through analysis of medical or economic data, to data analysis in terms of information systems security and detection of phenomena that may be dangerous. Both in one and the other case of enormous possibilities can be seen in the use of fuzzy logic, and in particular fuzzy logic systems. One of their main advantages is the fact that it is not necessary to specify "rigid" numerical values to determine "normal" values. It allows you to fuzzy this information and to "communicate" with the system in natural language. Classical methods require the definition of an acute value for the data area which is "normal". This is a great difficulty, especially when determining sharp values for data is difficult, and often impossible, e.g. searching for anomalies on X-ray images.

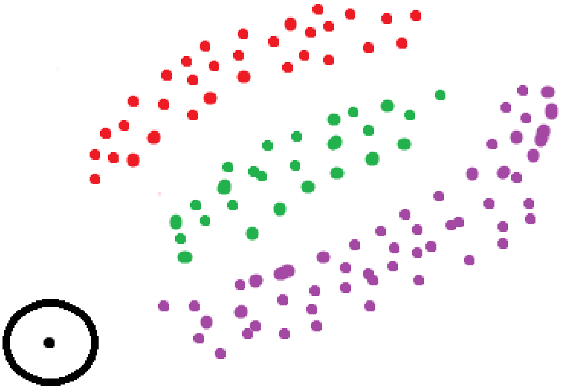

Figure 1: An example of a graphical representation of outlier

2. Preliminaries

Fuzzy set A in universe of discourse $\mathcal{X}$ is define as:

$$A = \{< x, \mu_A(x) >: x \in \mathcal{X}\} \tag{1}$$

where $\mu_A : \mathcal{X} \rightarrow [0, 1]$ is membership function and value $x \in \mathcal{X}$ is interpreted as membership degree of x to A.

Operations of the product, sum and complement for fuzzy sets A and B in $\mathcal{X}$ are defined in formulas (2) (3) and (4)

$$\mu_{A \cap B}(x) = min\{\mu_A(x), \mu_B(x)\} \tag{2}$$

$$\mu_{A \cup B}(x) = max\{\mu_A(x), \mu_B(x)\} \tag{3}$$

$$\mu_{A^c}(x) = 1 - \mu_A(x) \tag{4}$$

The *support* is set of all elements in universe of discourse $\mathcal{X}$ which membership degrees is greater then 0. It's can be defined as:

$$supp(A) = \{x \in \mathcal{X} : \mu_A(x) > 0\} \tag{5}$$

Interval type-2 fuzzy set in universe of discourse $\mathcal{X}$ is define as:

$$\widetilde{A} = \{< x, \underline{\mu}_{\widetilde{A}}(x), \overline{\mu}_{\widetilde{A}}(x) >: x \in \mathcal{X}\} \tag{6}$$

where $\underline{\mu}_{\widetilde{A}} : \mathcal{X} \to [0, 1]$ is lower membership function, $\overline{\mu}_{\widetilde{A}} : \mathcal{X} \to [0, 1]$ is upper membership function and $\widetilde{A}$ is:

$$\forall_{x \in \mathcal{X}}\ 0 \leq \underline{\mu}_{\widetilde{A}}(x) \leq \overline{\mu}_{\widetilde{A}}(x) \leq 1 \tag{7}$$

Extension of product (2), sum(3) and complement(4) take a form respectively:

$$\underline{\mu}_{\widetilde{A} \cap \widetilde{B}}(x) = min\{\underline{\mu}_{\widetilde{A}}(x), \overline{\mu}_{\widetilde{B}}(x)\},\ \overline{\mu}_{\widetilde{A} \cap \widetilde{B}}(x) = min\{\overline{\mu}_{\widetilde{A}}(x), \overline{\mu}_{\widetilde{B}}(x)\}, \tag{8}$$

$$\underline{\mu}_{\widetilde{A} \cup \widetilde{B}}(x) = max\{\underline{\mu}_{\widetilde{A}}(x), \underline{\mu}_{\widetilde{B}}(x)\},\ \overline{\mu}_{\widetilde{A} \cup \widetilde{B}}(x) = max\{\overline{\mu}_{\widetilde{A}}(x), \overline{\mu}_{\widetilde{B}}(x)\} \tag{9}$$

$$\underline{\mu}_{\widetilde{A}^c}(x) = 1 - \underline{\mu}_{\widetilde{A}}(x),\ \overline{\mu}_{\widetilde{A}^c}(x) = 1 - \overline{\mu}_{\widetilde{A}}(x) \tag{10}$$

Support of interval type-2 fuzzy set according to extend of (5) in universe $\mathcal{X}$ is define as:

$$\underline{supp}(\widetilde{A}) = \{x \in \mathcal{X} : \underline{\mu}_A(x) > 0\} \tag{11}$$

$$\overline{supp}(\widetilde{A}) = \{x \in \mathcal{X} : \overline{\mu}_A(x) > 0\} \tag{12}$$

Footprint Of Uncertainty (FOU) - is one of the most common ways to describe membership/belonging to interval type-2 fuzzy set. FUO is set of all pairs of $< x, u >$ on $X\ x\ J_X$ which value of secondary membership function is greater than 0 and in case of interval type-2 fuzzy sets value of secondary membership function is 1 Let $\widetilde{A}$ is interval type-2 fuzzy set in X, then footprint of uncertainty (FUO($\widetilde{A}$)) is defined as:

$$FUO(\widetilde{A}) = \{< x, u >: \mu_x(u) > 0, x \in X, u \in J_x\} \tag{13}$$

In interval type-2 fuzzy set can be define two specific elements. Lower membership function (LMF) and upper membership function (UMF). Those two function are primary membership functions and they define as:

$$LMF(\widetilde{A}) = \{< x, u >: x \in X, u = inf\ J_x\} \tag{14}$$

$$UMF(\widetilde{A}) = \{< x, u >: x \in X, u = sup\ J_x\} \tag{15}$$

$$\Sigma count(A) = \sum_{i=1}^{N} \mu_A(x_i) \tag{16}$$

Sigma-count, (generalized form cardinality) for classic fuzzy set A in finite universes of discourse $X = \{x_1, x_2, \dots, x_N\}, N \in \mathbb{N}$:

Relative cardinality of two fuzzy sets is defined as:

$$\Sigma count(A|B) = \frac{\Sigma count(A \cap B)}{\Sigma count(B)} \tag{17}$$

where $\Sigma count$ fuzzy set $A \cap B$ and B in $\mathcal{X}$ is count by (16)

Extension of Sigma-count for Interval Type-2 Fuzzy Sets is proposed by Niewiadomski [3]:

$$nf\sigma - count(\widetilde{A}) =_{df} \sum_{x \in X} sup\{u \in J_x : \mu_x(u) = 1\} \tag{18}$$

Relative cardinality of two Interval Type-2 Fuzzy Sets is defined as:

$$nf\sigma - count(\widetilde{A}|\widetilde{B}) =_{df} \frac{nf\sigma - count(\widetilde{A} \cap \widetilde{B})}{nf\sigma - count(\widetilde{B})} \tag{19}$$

where $\Sigma count$ fuzzy set $A \cap B$ and B in $\mathcal{X}$ is count by (16)

Wu and Mendel [4] define cardinality for interval type-2 fuzzy set as:

$$card_I(\widetilde{A}) =_{df} \frac{1}{2} \sum_{x \in X} ((LMF(\widetilde{A}))(x) + (UMF(\widetilde{A}))(x)) \tag{20}$$

A structure of type-2 fuzzy logic system is illustrated in Fig. 2.

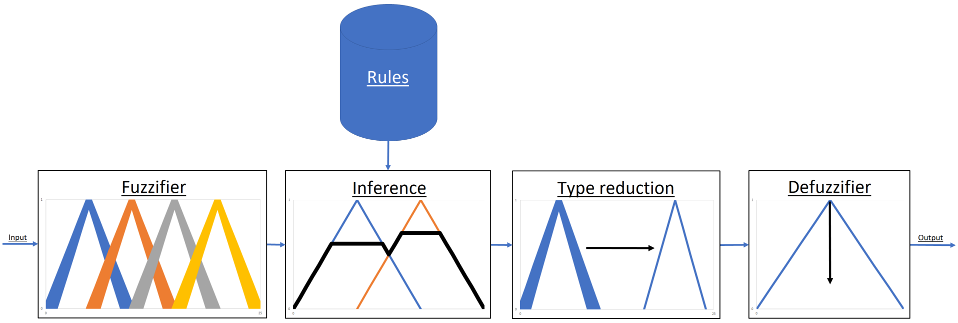

Figure 2: Type-2 Fuzzy Logic System: a general structure

3. Concept

The general concept of outliers search is to divide the space of considerations into areas described by fuzzy sets. Then it becomes possible to create fuzzy IF THEN rules. These rules, covering the whole space of considerations, constitute the basis for considering whether observation which deviates so much from the other to define it as being generated by another mechanism. Wu and Mendel in [5] they proposed a measure they referred to as Degree of Outlier, O (see eq.25). Although the article mainly focuses on linguistic summaries using the IF THEN rules, the proposed approach in the search for outliers in linguistic summaries is a strong basis for further research on the possibility of using fuzzy logic systems using the IF THEN rule base to detect and search for outliers . Two measures are used to determine which rules are rules that indicate the existence of an exception. First meter is The Degree of Truth (see eq.21) determining validity marked as T proposed by Kosko [6], it was also presented by van den Berg [7], Klir and Yuan as conditional and unqualified proposition [8] and in many other works to computing the confidence of fuzzy association rules [9, 10, 11]. The second one is The degree of sufficient coverage (see eq.22) which generally describes whether a given IF THEN rule is supported by sufficient data.

$$T = \frac{\sum_{m=1}^{M} min(\mu_{S_1}(v_1^m), \mu_{S_2}(v_2^m))}{\sum_{m=1}^{M} \mu_{S_1}(v_1^m)} \tag{21}$$

In general, it can specify that T is bigger when more data fulfills to IF THEN rule.

$$C = f(r_c) \tag{22}$$

where C is Degree of Sufficient Coverage, f is a function that maps r_c into C, and r_c is coverage ratio (see eq.23)

$$r_c = \frac{\sum_{m=1}^{M} t_m}{M} \tag{23}$$

where

$$t(m) = \begin{cases} 1, & \mu_{S_1}(v_1^m) > 0 \text{ and } \mu_{S_2}(v_2^m) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

Degree of Outlier defined as

$$O = \begin{cases} min(max(T, 1 - T), 1 - C), & T > 0 \\ 0, & T = 0 \end{cases} \tag{25}$$

allows you to convert rule values which, when specifying outlier, take values close to 0 or values close to 1. By using and properly defining functions for Degree of Sufficient Coverage, we are able to determine which data constitute the outlier in the set.

4. Conclusions

The use of fuzzy logic systems is in the search and discovery of outlier is a new area of research with very high research and implementation potential. Systems of this type not only allow you to search and find outlier, but above all they are a very natural way of communicating with people and determining the parameters of outlier search. To define individual parameters, no specialist knowledge is required, only a linguistic form is sufficient. In the case of outlier, which is often difficult to define in a mathematical way, it opens up new possibilities. Naturally, this area still requires a lot of research, but previous publications indicate great possibilities of the method.

5. Future work

Further research work will focus on creating universal methods for searching and detecting outliers, as well as new methods that will allow to more universalize the method as well as more accurately determine "degrees of truth" and "degrees of exception". It is also necessary to develop a universal method that allows the use of the cited and developed methods for any data, which will allow to solve the whole class of the problem, and not only specific cases. Despite many works indicating very good performance results of fuzzy logic systems, it is also necessary to research the performance of these systems in the search for outliers.

References

[1] Aggarwal, C. C. *Outlier Analysis*. Springer Science+Business Media New York, 2013.

[2] Hawkins, D. *Identification of Outliers*. Springer Netherlands, 1980.

[3] Niewiadomski, A. A type-2 fuzzy approach to linguistic summarization of data. *IEEE Transactions on Fuzzy Systems*, pages 198—-212, 2008.

[4] Wu, D. and Mendel, J. M. A vector similarity measure for interval type-2 fuzzy sets. *Proceedings of FUZZ-IEEE 2007 International Conference*, 2007.

[5] Wu, D., Mendel, J., and Joo, J. Linguistic summarization using if-then rules. pages 1 – 8. 2010. doi:10.1109/FUZZY.2010.5584500.

[6] Kosko, B. Fuzziness vs. probability. *International Journal of General Systems*, 17:11–240, 1990.

[7] van den Berg, J., Kaymak, U., and van den Bergh, W.-M. Fuzzy classification using probability-based rule weighting. *in Proc. IEEE Int'l Conf. on Fuzzy Systems*, pages 991–996, 2002.

[8] Klir, G. J. and Yuan, B. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Upper Saddle River, NJ:Prentice-Hall.

[9] Hong, T. P., Kuo, C. S., and Chi, S. C. Trade-off between computation time and number of rules for fuzzy mining from quantitative data. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(5):587–604, 2001.

[10] Ishibuchi, H., Nakashima, T., and Murata, T. Three-objective geneticsbased machine learning for linguistic rule extraction. *Information Sciences*, 136(1-4):109–133, 2001.

[11] Ishibuchi, H. and Yamamoto, T. Rule weight specification in fuzzy rulebased classification systems. *IEEE Trans. on Fuzzy Systems*, 13(4):428–435, 2005.